PCT/KR 2004/002117
RO/KR 23.08.2004

대한민국특허청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 :
Application Number 10-2003-0060528

출 원 년 월 일 :
Date of Application 2003년 08월 30일
AUG 30, 2003

출 원 인 :
Applicant(s) 주식회사 이즈텍
IStech CO., LTD.

2004 년 08 월 23 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】 특허출원서

【권리구분】 특허

【수신처】 특허청장

【제출일자】 2003.08.30

【발명의 명칭】 유전자 어휘 분류체계를 이용하여 바이오 칩을 분석하기 위한 시스템 및 그 방법

【발명의 영문명칭】 A SYSTEM FOR ANALYZING BIO CHIPS USING GENE ONTOLOGY, AND A METHOD THEREOF

【출원인】

【명칭】 주식회사 이즈텍

【출원인코드】 1-2000-030839-6

【대리인】

【성명】 김석현

【대리인코드】 9-1998-000634-1

【포괄위임등록번호】 2000-051369-8

【발명자】

【성명의 국문표기】 김양석

【성명의 영문표기】 KIM,Yang Suk

【주민등록번호】 680312-1093319

【우편번호】 411-380

【주소】 경기도 고양시 일산구 장항동 848-1 현대타운빌 506호

【국적】 KR

【발명자】

【성명의 국문표기】 허정욱

【성명의 영문표기】 HUR,Jung Uk

【주민등록번호】 750517-1807613

【우편번호】 745-886

【주소】 경상북도 문경시 점촌동 79-6

【국적】 KR

【발명자】

【성명의 국문표기】 이성근

【성명의 영문표기】 LEE,Sung Geun

【주민등록번호】 721215-1105916

1020[redacted]528

출력 일자: 2004/8/30

【우편번호】	608-090
【주소】	부산광역시 남구 용호동 LG 메트로시티 아파트 104동 1504호
【국적】	KR
【취지】	특허법 제42조의 규정에 의하여 위와 같이 출원합니다. 대리인 김석현 (인)
【수수료】	
【기본출원료】	20 면 29,000 원
【가산출원료】	17 면 17,000 원
【우선권주장료】	0 건 0 원
【심사청구료】	0 항 0 원
【합계】	46,000 원
【감면사유】	중소기업
【감면후 수수료】	23,000 원
【첨부서류】	1. 요약서·명세서(도면)_1통 2.중소기업기본법시행령 제2조에의한 중소기업에 해당함을 증명하는 서류[사업자등록증, 원천징수이행상황신고서]_1통

【요약서】

【요약】

본 발명은 유전자 어휘 분류체계(Gene Ontology; GO)의 계층 구조(hierarchical structure) 모델링을 통해 바이오 칩 또는 마이크로어레이 실험의 유전자 발현 양상(gene expression pattern)을 생물학적으로 분석하기 위한 시스템 및 그 분석 방법에 관한 것이다. 본 발명에 따른 GO를 이용한 바이오 칩 분석 시스템은 상기 바이오 칩 실험 결과의 통계적 클러스터링(clustering) 결과를 입력받아, 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 GO 용어 할당부; 상기 GO 용어 할당부에 의해 유전자에 할당된 GO 용어를 미리 설정된 숫자 조합인 GO 코드로 변환하는 GO 코드 변환부; 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하고, 계산된 유사거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하며, 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 용어 모두에 대해 상기 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 생물학적 의미 추출부를 포함한다.

【대표도】

도 1a

【색인어】

DNA 칩, 유전자, 어휘 분류체계(ontology), GO, GO 식별자, GO 코드

【명세서】

【발명의 명칭】

유전자 어휘 분류체계를 이용하여 바이오 칩을 분석하기 위한 시스템 및 그 방법 {A SYSTE FOR ANALYZING BIO CHIPS USING GENE ONTOLOGY, AND A METHOD THEREOF}

【도면의 간단한 설명】

도 1a는 GO 구조의 일례를 도시한 도면이고, 도 1b는 텍스트 구조의 GO의 일례를 도시한 도면.

도 2는 본 발명의 바람직한 일 실시예에 따른 GO를 이용한 DNA 칩 분석 시스템의 구성을 도시한 블록도.

도 3은 GO 용어를 GO 코드로 변환하는 일례를 설명하기 위한 도면.

도 4는 본 발명의 바람직한 일 실시예에 따른 생물학적 의미 추출부의 상세 구성을 도시한 블록도.

도 5는 GO 트리 구조의 두 노드 사이의 유사 거리를 구하는 일례를 도시한 도면.

도 6은 본 발명의 바람직한 일 실시예에 따른 GO를 이용한 DNA 칩 분석방법의 전체적인 흐름을 도시한 순서도.

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<7> 본 발명은 유전자 어휘 분류체계를 이용하여 바이오 칩을 분석하기 위한 시스템 및 그 방법에 관한 것으로서, 보다 구체적으로, 유전자 어휘 분류체계(Gene Ontology; 이하 'GO'라 한다) 계층 구조(hierarchical structure)의 모델링을 통해 DNA 칩 또는 마이크로어레이(Microarray) 실험의 유전자 발현 양상(gene expression pattern)을 생물학적으로 분석하기 위한 시스템 및 그 분석 방법에 관한 것이다.

<8> 1954년 와트슨 및 크릭(Watson and Crick)에 의하여 DNA의 이중 나선 구조가 밝혀진 이래 제한 효소의 발견, 혼성화(hybridization) 기법, PCR (Polymerase chain reaction) 등의 발전은 생명 현상의 분자 수준에서의 이해에 크게 기여하였다. 그러나 복잡한 조절 기능을 갖는 생명 현상을 단편적으로 이해하는 것이 아니라 인간 지놈 프로젝트(Human Genomic Project; HGP)와 같이 전체적 이해를 할 수 있는 실험의 필요성이 대두됨에 따라, 염기서열의 기능을 이해하기 위한 과정이 수행되는 가운데 DNA Chip이 개발되었다. 이러한 HGP와 DNA Chip의 결과를 효율적으로 활용하기 위하여 생물정보학(Bioinformatics)과 기능체 유전학(Functional Genomics)의 연구도 활발하게 진행되고 있다.

<9> 바이오 칩은 크게 마이크로어레이 및 마이크로플루이딕스(microfluidics) 칩으로 구분된다. 여기서 마이크로어레이는 수천개 혹은 수만개 이상의 DNA나 단백질 등을 일정 간격으로 배열하여 붙이고, 분석 대상 물질을 처리하여 그 결합 양상을 분석할 수 있는 칩을 말하며, DNA

칩 및 단백질 칩 등이 있다. 현재까지는 DNA 칩이 가장 널리 사용되고 있는 바이오 칩이라고 볼 수 있다. 또한, 마이크로플루이딕스 칩은 미량의 분석 대상 물질을 흘려보내면서 칩에 집적되어 있는 생물 분자 혹은 센서와 반응하는 양상을 분석하는데 이용된다.

<10> 이러한 DNA 칩은 유리판, 니트로셀룰로스 막(nitrocellulose membrane) 혹은 실리콘 위에 타겟 DNA 또는 cDNA나 올리고뉴클레오티드(oligonucleotide)를 붙인 것이다. 다시 말하면, 이러한 DNA 칩은 작은 면적의 고체 표면에 염기서열이 알려진 cDNA 혹은 올리고뉴클레오티드 탐침(probe)을 정해진 위치에 미세 집적(micro-array)시킨 것을 말한다.

<11> 이러한 DNA 칩은 형광물질 혹은 방사선 동위 원소로 표식된 탐침과 혼성화시켜 유전자의 발현 정도, 돌연 변이의 확인, 단일 뉴클레오티드 다형성(single nucleotide polymorphism; SNP), 질병의 진단, 고처리 스크리닝(high-throughput screening; HTS) 등에 사용할 수 있다. 이러한 DNA 칩에 분석하고자 하는 시료 DNA 단편을 결합시키면, DNA 칩에 부착되어 있는 탐침과 시료 DNA 단편상의 염기서열의 상보적 정도에 따라 혼성화 상태를 이루게 되는데, 광학적인 방법 혹은 방사능 화학적 방법 등을 통해 이를 관찰 해석함으로써, 시료 DNA의 염기 서열을 측정할 수 있다. 이러한 DNA 칩을 이용하면 많은 수의 유전자의 발현 정보를 간편하고 신속하게 알 수 있으며, 현재 신약 개발 및 의료 진단용으로 개발 사용되고 있다.

<12> DNA 칩 결과의 분석에는 통계적인 방법과 생물학적인 방법이 병행되고 있다. 이미지 분석을 통하여 나타난 각 유전자들의 발현 정도를 통계적인 방법을 이용하여 공통적인 발현 양상을 보이는 것들을 클러스터링(clustering)을 통하여 묶어 낸다.

<13> 클러스터는 통계적인 방법에 의해서만 생성된 것으로서, 이에 대한 생물학적인 확인을 위해, 클러스터에 포함된 각 유전자의 알려진 기능을 이용하여 해당 클러스터(cluster)에 일반적인 의미를 부여함과 동시에 해당 클러스터의 신뢰도를 생물학적으로 확인하게 된다.

<14> 기존의 생물학적 확인 과정은 논문이나 기존의 생물학 정보 데이터베이스 등에서 유전자의 기능을 추출하여 비교하는 방법을 이용한다. 이때 사용되는 데이터베이스들은 NCBI(National Center for Biotechnology Information)의 기본적인 DNA 정보, MIPS(Munich information center for protein sequences) 혹은 CGAP(Cancer genome anatomy project) 등의 기능별 분류(functional category) 정보, 또는 Swiss-Prot의 단백질 정보들을 이용한다.

<15> 그러나, 위와 같은 클러스터의 생물학적 의미를 판단하기 위한 작업은 수작업을 통해서 많이 이루어지고 있으며, 생물학 용어의 다양성 등으로 인하여 체계적이고 자동화된 분석을 수행하기 어려웠다는 문제점이 있다.

<16> 또한, 기존 생물학 정보 데이터베이스의 경우, 단백질의 정보원으로 많이 사용되는 Swiss-Prot은 핵심 단어(keyword)를 이용하여 단백질들의 기능을 잘 분류하였으나, 이들 핵심 단어들 사이에는 정형화된 상관 관계 혹은 상하 관계(hierarchy)가 존재하지 않으며, 이 때문에 DNA 칩의 생물학적 분석에서 자동화에 장애 요인으로 작용한다. 또한, CGAP(Cancer Genome Anatomy Project) 등의 특화된 분야별의 그룹 정보들은 해당 분야에서만 적용되는 한계점을 지니며, 또한 그 그룹 자체가 너무 넓은 의미의 기능을 다루게 되므로, 세부적인 기능적 측면에서는 한계점을 지니게 된다는 문제점이 있다.

<17> 따라서, 종래에는 통계적인 방법에 의해서만 생성된 클러스터에 생물학적 의미를 부여하는데는 오랜 시간이 걸릴 수밖에 없었으며, 아울러, 세부적이고 정확한 생물학적 의미를 부여하기 어려운 문제점이 있었다.

<18> 한편, GO 컨소시엄(Gene Ontology Consortium)에서는 GO 용어를 제공하고 있는데, 여기서 어휘 분류체계(Ontology)란 간략하게 말하면 생물학 용어 또는 어휘를 분류해 놓은 체계를 말한다. 유전자 어휘 분류체계 컨소시엄은 생물학 용어들의 통합을 목적으로 세워졌으며, 모

든 생물 종들에서 유전자의 기능을 설명하는데 있어서 사용되는 공통적으로 사용될 수 있는 통합된 용어들을 제공하며, 현재 일만여개의 용어로 구성되어 있다. 결국, GO는 유전자(Gene) 혹은 유전자에 함축된 키워드들이 각 개체가 되어 그것들 사이의 관계를 연구하는 것을 의미하며, 생물정보학(bioinformatics)에 적용하게 된다.

<19> 이러한 GO 용어의 특이점은 각 용어들 사이에 상하 관계의 트리 구조를 가지며, 전체 용어들을 3가지의 큰 범주(category)로 구분된다는 점이다. 즉, 세개의 큰 범주를 가지고 약 10,000개 정도의 용어들이 마치 트리 구조처럼 상하 관계(hierarchy)를 가지고 구성이 되어 있다. 이것을 이용하여 DNA 칩의 분석시 생물학적 의미를 찾기 위한 것으로, GO는 유전자의 기능을 크게 i) 분자의 기능(molecular function), ii) 생물학적 작용(biological process), 및 iii) 세포 성분(cellular component)의 범주로 나누고, 각각의 범주에 계층적인 통제 어휘(controlled vocabulary)를 확립하였다. 이들 범주는 서로 배타적인 것이 아니며, 한 개의 유전자를 묘사하기 위한 특징들을 나누는 범주이다.

<20> 본 발명은 이러한 GO 용어들을 이용하여 클러스터에 자동화된 방법으로 생물학적 의미를 부여하는 시스템 및 방법에 관한 것이다.

【발명이 이루고자 하는 기술적 과제】

<21> 전술한 문제점을 해결하기 위한 본 발명의 목적은 GO 계층 구조의 모델링을 통해 DNA 칩 실험의 유전자 발현 양상에 대해 체계적으로 생물학적 분석을 수행할 수 있도록 유전자 어휘 분류체계를 이용하여 바이오 칩을 분석하기 위한 시스템 및 분석 방법을 제공하기 위한 것이다

<22> 또한, 본 발명의 다른 목적은 GO 용어와 트리 구조를 이용하여 바이오 칩의 실험 결과의 통계적인 클러스터링(clustering)을 통해 생성되는 클러스터(cluster)에 속하는 유전자들의 가장 공통적이며 이상적인 유전자의 기능을 추출하는 방법을 제공하기 위한 것이다.

【발명의 구성】

<23> 상기한 바와 같은 목적을 달성하기 위하여, 본 발명의 바람직한 일 실시예에 따르면, 상기 바이오 칩 실험 결과의 통계적 클러스터링(clustering) 결과를 입력받아, 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 GO 용어 할당부; 상기 GO 용어 할당부에 의해 유전자에 할당된 GO 용어를 미리 설정된 숫자 조합인 GO 코드로 변환하는 GO 코드 변환부; 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하고, 계산된 유사거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하며, 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 용어 모두에 대해 상기 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 생물학적 의미 추출부를 포함하는 바이오 칩 분석 시스템이 제공된다.

<24> 상기 GO 용어 할당부는 생물학 데이터베이스를 마이닝한 결과로부터 유전자에 상응하는 GO 용어를 할당할 수 있다.

<25> 상기 GO 코드 변환부는 변환하려는 GO 용어가 속하는 레벨, 변환하려는 GO 용어의 모(母)노드 정보 및 변환하려는 GO 용어가 속하는 레벨에서의 순서에 따라 GO 용어를 GO 코드로 변환할 수 있다.

<26> 상기 생물학적 의미 추출부는,

<27> 상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 최적 교차점들을 추출하는 최적 교차점 추출부;

<28> 상기 최적 교차점 정보를 이용하여 상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 유사 거리들을 계산하는 유사 거리 계산부;

<29> 상기 유사 거리 계산부에서 계산된 유사 거리들의 평균 유사 거리를 계산하는 평균 유사 거리 계산부;

<30> 상기 유사 거리 계산부에서 계산된 유사 거리들 중 최대 유사 거리를 판단하는 최대 유사 거리 판단부;

<31> 상기 미리 설정된 그룹에 속하는 모든 GO 용어에 대한 평균 유사 거리 및 최대 유사 거리를 비교하여, 최소의 평균 유사 거리 또는 최소의 최대 유사 거리에 상응하는 GO 트리 구조상의 용어를 상기 클러스터에 대응하는 최적의 매칭 노드로 판단하는 최적 매칭 노드 판단부를 포함할 수 있다.

<32> 상기 미리 설정된 그룹의 속하는 GO 용어는 GO 트리 구조에 포함된 모든 용어일 수 있다

<33> 상기 미리 설정된 그룹의 속하는 GO 용어는 사용자가 선택한 GO 트리 구조 레벨에 상응하는 GO 용어일 수 있다.

<34> 상기 최적 교차점 추출부는 GO 트리 구조에서 두 개의 GO 용어를 모두 하위 레벨에 포함하는 상위 GO 용어들 중 가장 하위 레벨에 속하는 GO 용어를 최적 교차점으로 판단할 수 있다.

<35> GO 트리 구조의 각 레벨에는 미리 설정된 가중치가 부여되어 있으며, 상기 유사 거리 계산부에서 계산되는 유사 거리는 두 GO 용어의 최적 교차점이 속하는 레벨의 가중치일 수 있다.

<36> 한편, 본 발명의 다른 실시예에 따르면, a) 상기 바이오 칩 실험 결과의 통계적 클러스터링 결과를 입력받아 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 단계; b) 상기 유전자마다 할당된 GO 용어를 각각 미리 설정된 숫자 조합인 GO 코드로 변환하는 단계; c) 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하는 단계; d) 상기 단계(c)에서 계산된 유사 거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하는 단계; 및 e) 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 모두에 대해 상기 단계 (c) 및 (d)를 반복하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 단계를 포함하는 것을 특징으로 하는 바이오 칩 분석 방법이 제공된다.

<37> 한편, 본 발명의 또 다른 실시예에 따르면, 바이오 칩 분석 방법을 실행하기 위하여 디지털 처리 장치에 의해 실행될 수 있는 명령어들의 프로그램이 유형적으로 구현되어 있으며, 디지털 처리 장치에 의해 판독할 수 있는 기록 매체에 있어서, 상기 바이오 칩 분석 방법은, a) 상기 바이오 칩 실험 결과의 통계적 클러스터링 결과를 입력받아 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 단계; b) 상기 유전자마다 할당된 GO 용어를 각각 미리 설정된 숫자 조합인 GO 코드로 변환하는 단계; c) 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하는 단계; d) 상기 단계(c)에서 계산된 유사 거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하는 단계; 및 e) 상기 미리 설정된 그

룹에 속하는 GO 트리 구조상의 GO 용어 모두에 대해 상기 단계 (c) 및 (d)를 반복하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 단계를 포함하는 기록 매체가 제공된다.

<38> 이하 첨부된 도면을 참조하여 본 발명에 따른 GO를 이용하여 DNA 칩을 분석하기 위한 시스템과 그 방법의 바람직한 실시예를 설명한다.

<39> 도 1a는 GO 구조의 일례를 도시한 도면이고, 도 1b는 텍스트 구조의 GO의 일례를 도시한 도면이다.

<40> 본 발명에 대한 설명에 앞서, GO의 계층 구조에 대해 살펴보기로 한다. 도 1a에 도시된 바와 같이, GO 계층 구조에서 최상위 레벨은 GO 계층이고, 두 번째 계층은 전술한 분자의 기능(molecular function), 생물학적 작용(biological process), 및 세포 성분(cellular component) 계층에 해당하며, 레벨 3, 4 및 5의 하위 레벨로 각각 트리가 형성되며, 하위 레벨로 갈수록 더 세부적인 기능의 GO 용어가 존재한다. 도 1에 도시된 바와 같이, GO 구조는 완벽한 트리 구조가 아닌 유향 그래프 구조이다. 본 발명에서는 도 1에 도시된 GO 구조가 아닌 유향 그래프 구조의 GO 구조를 트리 구조로 변환한 모델이 사용된다. 유향 그래프 구조를 트리 구조로 변환하는 방법은 단순하며 이미 공지된 것이기에 이에 대한 상세한 설명은 생략하기로 한다. 도 1b는 트리 구조로 변환된 GO 모델을 다시 텍스트 형식으로 나타낸 것으로서, 하위 레벨의 GO 용어는 상위 레벨의 GO 용어보다 오른쪽으로 치우친 열에 기록되며, 같은 레벨의 GO 용어는 같은 열에 기록된다. 텍스트 구조의 GO 모델은 GO 컨소시엄으로부터 제공받을 수 있다.

<41> 도 2는 본 발명의 바람직한 일 실시예에 따른 GO를 이용한 DNA 칩 분석 시스템의 구성을 도시한 블록도이다.

<42> 도 2에 도시된 바와 같이, 본 발명의 일 실시예에 따른 DNA 칩 분석 시스템은 클러스터링부(200), GO 용어 할당부(202), GO 코드 변환부(204), GO 코드 저장부(206) 및 생물학적 의미 추출부(208)를 포함할 수 있다.

<43> 클러스터링부(100)는 DNA 칩의 발현량 데이터를 이용하여 발현 패턴이 유사한 유전자들에 대한 클러스터링을 수행한다. DNA 칩의 발현량은 다양한 조건하에서 구해지며, 클러스터링은 DNA 칩에 포함된 복수의 유전자 중 발현 패턴이 유사한 유전자들을 하나의 군으로 묶는 과정을 의미한다. 따라서, 클러스터링 결과 복수의 클러스터가 생성될 수 있으며, 각 클러스터에는 발현패턴이 유사한 복수의 유전자들이 포함된다. 클러스터링에 대해서는 다양한 알고리즘이 공지되어 있으므로, 이에 대한 상세한 설명은 생략하기로 하며, 공지된 다양한 형태의 클러스티링 알고리즘이 본 발명에 적용될 수 있을 것이다.

<44> GO 용어 할당부(202)는 클러스터링 수행 후 클러스터에 포함된 각각의 유전자들에 대하여 관련된 GO 용어를 할당하는 기능을 한다. 이는 클러스터에 포함된 어떠한 유전자의 기능이 GO에서 정의하는 기능들 중 어떠한 용어에 해당되는지를 판단하고, 해당 GO 용어를 유전자에 할당하는 것을 의미한다. 유전자가 복수의 기능을 수행할 경우, 복수의 GO 용어가 유전자에 할당될 수 있을 것이다.

<45> 본 발명의 일 실시예에 따르면, 네트워크를 통해 생물학 데이터베이스로부터 특정 유전자와 관련되는 GO 용어를 획득할 수 있을 것이다. 네트워크를 통해 액세스할 수 있는 생물학 데이터베이스에는 Unigene, LocusLink, Swiss-Prot, MGI 등이 있으며, 이에 한정되지는 않는다. 위의 데이터베이스들 중 대부분은 유전자의 기능과 관련된 GO 용어를 제공하고 있으며, 유전

자와 관련된 GO 용어를 제공하지 않더라도 데이터베이스에서 제공하는 유전자의 기능 정보를 이용하여 해당 GO 용어를 찾아낼 수 있을 것이다. 여기서 UniGene은 NCBI(National Center for Biotechnology Information)에서 제공하는 DNA 수준에서의 유전자 정보를 제공하고, LocusLink는 NCBI의 대표 서열 프로젝트(Reference Sequence Project)의 결과로 각 유전자별 기능 및 대표성을 가지는 서열 정보를 제공하며, Swiss-Prot은 스위스 생물정보학 연구소(Swiss Institute of Bioinformatics)에서 단백질 수준의 정보를 제공하며, 그리고 MGI는 쥐의 유전체 정보를 제공한다.

<46> 본 발명의 다른 실시예에 따르면, 위와 같이 네트워크를 통해 액세스 할 수 있는 데이터베이스 이외에, 자체적으로 구축한 데이터베이스 또는 파일을 통해 유전자에 상응하는 GO 용어를 할당할 수도 있을 것이다.

<47> GO 코드 변환부(204)는 상기 유전자에 할당된 GO 용어를 미리 설정된 GO 코드로 변환하는 기능을 한다. GO 용어는 문자이므로 다른 GO 용어들과 GO 트리 구조상에서 어느 정도 근접해있는지 여부를 판단할 수 없다. 따라서, 본 발명에서는 GO 용어를 미리 설정된 숫자 조합으로 변환하도록 한다. GO 용어를 숫자들의 조합으로 변환함으로써, GO 트리 구조상에서 특정 노드의 GO 용어가 다른 노드의 GO 용어와 어느 정도의 밀접한 관계가 있는지를 수치상으로 계산할 수 있게 된다.

<48> GO 용어를 GO 코드로 변환하는 구체적인 방법 및 GO 코드의 구체적인 구성은 별도의 도면을 통해 후술하기로 한다.

<49> GO 코드 저장부(206)는 GO 트리 구조의 GO 용어들을 미리 GO 코드로 변환한 정보들을 저장하고 있으며, GO 코드 변환부(204)는 GO 코드 저장부(206)에 저장된 정보를 이용하여 GO 용어를 GO 코드로 변환할 수 있을 것이다.

<50> 생물학적 의미 추출부(208)는 유사한 발현 패턴을 가지는 유전자들의 집합인 클러스터가 생물학적으로 어떠한 의미를 가지는가를 판단하는 기능을 한다. 생물학적 의미 추출부(208)는 클러스터에 포함된 유전자들의 기능이 GO 트리 구조에 포함된 용어 중 어떠한 GO 용어에 가장 근접하는지 여부를 판단하고, 가장 근접하는 GO 용어를 해당 클러스터에 연관시킴으로써 클러스터에 포함된 유전자들의 대표적인 기능을 판단할 수 있도록 한다.

<51> 전술한 바와 같이, 클러스터링은 생물학적인 의미와는 관계없이 통계적인 방법에 의해서만 이루어지므로 하나의 클러스터에 대해 생물학적인 의미를 찾는데 많은 시간이 소요되었다. 본 발명에 따르면, 클러스터가 어떠한 GO 용어에 가장 가까운가를 프로그램에 의해 미리 판단함으로써 클러스터의 생물학적인 분석 작업시간을 현저히 줄일 수 있게 된다.

<52> 클러스터가 어떠한 GO 용어에 가장 근접하는지를 판단하기 위해, 생물학적 의미 추출부(208)는 GO 트리 구조상의 한 노드와 클러스터에 포함된 각각의 유전자들과의 근접도를 계산한다. 근접도의 계산을 위해 본 발명에서는 유사 거리(Pseudo Distance)라는 개념을 도입하며, 유사 거리를 계산하는 방법은 후에 상세히 설명한다.

<53> 생물학적 의미 추출부(208)는 GO 트리 구조상의 한 노드와 클러스터에 포함된 모든 유전자들과의 유사 거리를 계산한 후, GO 트리 구조상의 한 노드와 클러스터에 포함된 유전자들과의 평균 유사 거리 및 최대 유사 거리를 계산한다.

<54> 상술한 GO 트리 구조상의 한 노드와 클러스터에 포함된 유전자들 사이의 평균 유사 거리 및 최대 유사 거리를 계산하는 과정은 GO 트리 구조상의 모든 노드 또는 선택된 일부의 노드에 대해 이루어질 수 있고, 이중 가장 짧은 평균 유사 거리를 가지는 GO 트리 구조의 노드 및 가장 짧은 최대 유사 거리를 가지는 GO 트리 구조의 노드를 클러스터와 가장 근접한 노드로 판단하며, 해당 노드의 GO 용어를 클러스터의 생물학적인 의미로 판단할 수 있을 것이다.

<55> 도 3은 GO 용어를 GO 코드로 변환하는 일례를 설명하기 위한 도면이다.

<56> GO 용어는 GO 트리 구조에서 GO 용어가 속하는 레벨 및 레벨에서의 순서에 따라 GO 코드로 변환된다.

<57> 도 3에서, 식별부호 300의 GO 용어는 1레벨에 속하며, 1레벨의 첫 번째 노드이다. 이때 식별부호 300의 GO 용어는 "10000000000000"의 GO 코드로 변환된다. GO 코드가 15자리인 것은 현재의 GO 레벨이 15레벨이기 때문이며, GO 코드의 첫 번째 자리는 1레벨에서의 값, GO 코드의 두 번째 자리는 2레벨에서의 값을 각각 나타낸다. 식별부호 300의 GO 용어는 1레벨의 첫 번째 GO 용어이므로 2번째 자리수부터 15번째 자리수까지의 값은 0이고, 첫 번째 자리수의 값은 1이다.

<58> 식별부호 302의 GO 용어는 2번째 레벨이며, 식별 부호 300인 GO 용어의 하위 노드이다. 이때, 식별부호 302의 GO 용어는 "110000000000000"의 GO 코드로 변환된다.

<59> 식별부호 302의 GO 용어는 2레벨에 속하기 때문에, 3자리부터 15자리까지의 값은 0이다. 또한, 식별부호 300에 해당하는 GO 노드의 자(子)노드이기 때문에, 첫 번째 자리수의 값은 모(母)노드의 값을 그대로 사용한다. 또한, 식별부호 302의 GO 용어는 레벨2에 속하는 식별부호 300의 노드의 하위 노드들 중 첫 번째 노드이am로 2번째 자리수의 값은 1이다.

<60> 이와 같은 원리로, 식별 부호 304의 GO 용어는 "120000000000000"의 GO 코드로 변환될 수 있을 것이다.

<61> 식별 부호 310의 GO 용어는 세 번째 레벨이고, 식별 부호 302의 노드의 자(子)노드이며, 식별 부호 302의 자(子)노드들 중 2번째 노드이다. 따라서, 식별 부호 310의 GO 용어는 "

11200000000"의 GO 코드로 변환될 수 있을 것이다. 같은 원리로, 식별부호 312의 GO 용어는 " 121000000"의 GO 코드로 변환된다.

<62> 위와 같은 원리로 GO 용어가 GO 코드로 변환되므로, GO 코드는 GO 용어가 속하는 레벨 및 GO 용어의 모(母)노드에 대한 정보를 포함하고 있다.

<63> 도 4는 본 발명의 바람직한 일 실시예에 따른 생물학적 의미 추출부의 상세 구성을 도시한 블록도이다.

<64> 도 4에 도시된 바와 같이, 본 발명의 일 실시예에 따른 생물학적 의미 추출부는 최적 교차점 추출부(400), 유사 거리 계산부(402), 평균 유사거리 계산부(404), 최대 유사 거리 판단부(406) 및 최적 매칭 노드 판단부(408)를 포함할 수 있다.

<65> 최적 교차점 추출부(400)는 유사 거리를 계산하기 위한 두 개의 노드 사이의 최적 교차점을 추출하는 기능을 한다. 최적 교차점의 추출은 유사 거리를 구하기 위한 전 단계로서, 두 개의 노드 사이의 최적 교차점이란 GO 트리 구조상에서 두 개의 노드를 모두 아래에 포함하는 상위 노드들 중 가장 하위 레벨에 속하는 노드를 의미한다.

<66> 예를 들어, 도 3을 참조하면, 식별 부호 308의 노드와 식별 부호 310의 노드를 모두 포함하는 상위 노드는 식별 부호 302의 노드 및 식별 부호 300의 노드가 있다. 이중 식별 부호 302의 노드가 가장 하위 노드이므로, 식별부호 308의 노드 및 식별 부호 310의 노드의 최적 교차점은 식별부호 302의 노드이다.

<67> GO 코드를 이용할 경우, 최적 교차점은 비교적 쉽게 구해질 수 있다. 도 3에서, 식별부호 308번 노드의 GO 코드는 "111000000000000"이고, 식별 부호 310번 노드의 GO 코드는 " 112000000000000"이다. 두 개의 GO 코드는 2번째 자리까지 동일하므로, 최적 교차점은 2번째

레벨에 존재하며, 1레벨의 첫 번째 노드(첫번째 자리수가 1이므로)의 자(子)노드들 중 첫 번째 노드(두번째 자리수가 1)가 최적 교차점이라는 것을 알 수 있다.

<68> 유사 거리 계산부(402)는 상기 최적 교차점 정보를 이용하여 GO 트리 구조상에서 두 노드 사이의 유사 거리를 계산하는 기능을 한다. 전술한 바와 같이, GO 트리 구조의 특정한 한 GO 용어(노드)와 클러스터에 포함된 모든 유전자에게 각각 할당된 GO 용어(노드)사이의 유사거리가 계산되며, 이와 같은 유시 거리 계산은 GO 트리 구조상의 모든 노드 또는 선택된 일부에 대하여 수행된다.

<69> 본 발명의 일 실시예에 따르면, GO 트리 구조의 각 레벨에는 가중치(Weight)가 부여되며, 두 노드 사이의 유사 거리는 두 노드의 최적 교차점이 속하는 레벨의 가중치로 정의할 수 있다. 단 두 개의 노드가 동일할 경우에 유사 거리는 0으로 정의된다.

<70> 도 5는 GO 트리 구조의 두 노드 사이의 유사 거리를 구하는 일례를 도시한 도면이다.

<71> 도 5에 도시된 바와 같이, GO 트리 구조의 각 레벨에는 가중치가 부여되어 있다(1레벨-150, 2레벨 140 등). 도 5에서, 식별 부호 500의 노드와 식별 부호 502의 노드의 최적 교차점은 식별 부호 504 노드이다. 식별 부호 504 노드는 3레벨에 존재하며, 3레벨에 부여된 가중치는 130이다. 따라서, 식별 부호 500의 노드와 식별 부호 502의 노드의 유사 거리는 130으로 계산될 수 있다.

<72> 평균 유사 거리 계산부(404)는 유사 거리 계산부(402)에서 GO 트리 구조상의 특정한 GO 용어와 하나의 클러스터에 포함된 모든 유전자에 할당된 GO 용어들 사이의 유사 거리가 계산된 후, 계산된 유사 거리들의 평균을 구하는 기능을 한다. 계산된 평균 유사 거리는 GO 트리 구조상의 특정한 노드와 클러스터 사이의 관련도를 나타내는 척도로 사용된다.

<73> 최대 유사 거리 판단부(406)는 유사 거리 계산부(402)에서 GO 트리 구조상의 특정한 GO 용어와 클러스터에 포함된 모든 유전자에 할당된 GO 용어들 사이의 유사 거리가 계산된 후, 계산된 유사 거리들 중 최대값을 추출하는 기능을 한다. 최대 유사 거리가 클수록 해당 클러스터는 소속 유전자의 일반적인 공통성을 해치는 부적당(bad)한 유전자를 포함하고 있을 가능성이 높게 된다. 클러스터는 수학적인 방법으로 발현 패턴이 유사한 유전자들을 모아놓은 집합으로, 생물학적인 공통성이 충분히 고려된 것은 아닌 바, 최대 유사 거리를 계산함으로써 소속 유전자들의 생물학적인 공통성을 판단할 수 있게 된다.

<74> 최적 매칭 노드 판단부(408)는 GO 트리 구조상의 모든 노드에 대해 상기 클러스터와의 평균 유사 거리 및 최대 유사 거리가 계산된 후 가장 작은 평균 유사 거리를 가지는 노드 및 가장 작은 최대 유사 거리를 가지는 노드를 판단하고, 이 노드를 해당 클러스터와 최적으로 매칭이 되는 노드라 판단한다. 따라서, 판단된 노드에 상응하는 GO 용어가 해당 클러스터를 대표하는 용어가 되며, 통계적인 방법으로 형성된 클러스터에 생물학적인 의미를 부여할 수 있게 된다. 가장 작은 평균 유사 거리를 가지는 노드 및 가장 작은 최대 유사 거리를 가지는 노드는 동일할 수도 있으며, 그렇지 않을 수도 있다. 또한, 최적 매칭 노드 판단부(408)는 가장 작은 평균 유사 거리 정보 및 가장 작은 최대 유사 거리 정보 중 하나만을 이용하여 최적으로 매칭이 되는 노드를 판단할 수도 있을 것이다.

<75> 도 6은 본 발명의 바람직한 일 실시예에 따른 GO를 이용한 DNA 칩 분석방법의 전체적인 흐름을 도시한 순서도이다.

<76> 도 6에 도시된 바와 같이, 본 발명에 따른 방법은 DNA 칩 실험 결과의 통계적 클러스터링(clustering) 결과를 입력받는 단계(S10), 각 클러스터에 속하는 유전자마다 Gene Ontology(GO) 용어를 할당하는 단계(S20); GO 코드 파일을 이용하여 상기 유전자마다 할당된

GO 용어를 각각 GO 코드로 변환하는 단계(S30); 변환된 GO 코드를 이용하여 GO 트리 구조상의 특정 노드와 클러스터에 포함된 모든 유전자에 할당된 GO 노드들 사이의 유사 거리를 계산하는 단계(S40); 상기 S40 단계에서 계산한 유사 거리들의 평균 유사 거리를 구하는 단계(S50); 상기 S40단계에서 계산한 유사 거리들의 최대 유사 거리를 구하는 단계(S60); 및 GO 트리 구조상의 모든 노드에 대해 클러스터와의 평균 유사 거리 및 최대 유사 거리를 계산하여(S70), 가장 작은 평균 유사 거리를 가지는 노드 및 가장 작은 최대 유사 거리를 가지는 노드를 클러스터에 연관시킴으로써 클러스터의 생물학적 의미를 추출하는 단계(S80)를 포함한다.

<77> 도 6을 참조하여, 본 발명에 따른 GO 구조를 이용한 DNA 칩의 유전자 발현 양상의 생물학적 분석 방법을 상세히 설명하면 다음과 같다.

<78> 먼저, 유전자 발현 양상의 통계적 클러스터링 결과로부터 각 클러스터에 속하는 유전자별로 GO 용어를 할당하고, 할당된 GO 용어를 GO 코드로 변환하는 과정을 수행하게 된다.

<79> 구체적으로, 클러스터링 결과를 입력(S10)하면, 각 유전자에 상응하는 GO 용어를 여러 데이터베이스의 마이닝(mining)을 통해 획득하고, 획득한 GO 용어를 해당 유전자에 할당한다(S20). 이때, 데이터베이스 마이닝을 통해 GO 용어를 미리 할당해 놓은 파일을 이용하여, 클러스터내의 유전자들에 GO 용어를 할당할 수도 있을 것이다. 다음에, GO 트리 구조 전체를 코드화 시켜놓은 GO 코드 파일을 이용하여, 클러스터의 유전자에 할당한 GO 용어를 GO 코드로 변환하게 된다(S30).

<80> GO 코드로의 변환 후, GO 트리 구조의 특정 노드와 클러스터에 포함된 모든 유전자에 할당된 GO 용어(노드) 사이의 유사 거리가 계산된다(S40). 전술한 바와 같이, 두 노드 사이의 유사 거리 계산을 위해 최적 교차점이 추출되며, 추출된 최적 교차점이 속하는 레벨의 가중치를 유사 거리로 판단한다.

<81> GO 트리 구조의 특정 노드와 클러스터에 포함된 유전자에 할당된 GO 용어(노드) 사이의 유사 거리가 계산된 후, 계산된 유사 거리들의 평균값을 구하고(S50), 계산된 유사 거리들 중 최대 값을 구한다(S60).

<82> 상기 특정 GO 노드와 클러스터의 포함된 유전자들 사이의 유사 거리를 계산하는 과정은 GO 트리 구조의 모든 노드에 대해 이루어진다(S70). 이때, 클러스터와의 평균 유사 거리가 가장 작은 GO 노드 및 클러스터와의 최대 유사 거리가 가장 작은 GO 노드를 해당 클러스터와의 최적 매칭 노드로 판단하고, 해당 GO 노드에 상응하는 GO 용어를 클러스터를 대표하는 생물학적 기능이라고 판단한다(S80). 여기서, 가장 작은 평균 유사 거리를 가지는 GO 노드 및 가장 작은 최대 유사 거리를 가지는 GO 노드 모두가 최적 매칭 노드 판단에 반드시 이용되어야 하는 것은 아니며, 이중 하나의 노드만이 최적 매칭 노드 판단에 이용될 수 있다는 것은 당업자에게 있어 자명할 것이다.

<83> 본 발명의 다른 실시예에 따르면, GO 트리 구조의 모든 노드에 대해 클러스터와의 평균 유사 거리가 계산되지 않고, 사용자가 선택한 특정 레벨에 포함된 노드에 대해서만 클러스터와의 평균 유사 거리가 계산될 수도 있으며, 이 경우 사용자가 선택한 특정 레벨에 포함된 GO 용어 중 하나가 클러스터의 생물학적 의미로 부여될 수 있을 것이다. 미리 레벨을 지정하여 생물학적 의미를 추출할 경우, 비교적 알기 힘든 하위 레벨에서의 생물학적 의미도 쉽게 유추될 수 있을 것이다.

<84> 상기의 실시예는 DNA 칩의 분석 방법에 대하여 기술되었으나, 단백질 칩 등을 포함하는 다른 바이오 칩에 대해서도 본 발명이 적용될 수 있다는 것은 당업자에게 있어 자명할 것이다.

<85> 본 발명을 상기 실시예에 의해 구체적으로 설명하였지만, 본 발명은 이에 의해 제한되는 것은 아니고, 당업자의 통상적인 지식의 범위 내에서 그 변형이나 개량이 가능하다.

【발명의 효과】

<86> 본 발명에 따르면, GO 계층 구조의 모델링을 통해 DNA 칩 실험의 유전자 발현 양상에 대해 체계적으로 자동화된 생물학적 분석을 수행할 수 있고, 또한 GO 용어와 트리 구조를 이용하여 DNA 칩의 실험 결과의 통계적인 클러스터링을 통해 생성되는 클러스터에 속하는 유전자들의 가장 공통적이며 이상적인 유전자의 기능을 추출할 수 있다.

【특허청구범위】

【청구항 1】

바이오 칩을 분석하기 위한 시스템에 있어서,

상기 바이오 칩 실험 결과의 통계적 클러스터링(clustering) 결과를 입력받아, 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 GO 용어 할당부;

상기 GO 용어 할당부에 의해 유전자에 할당된 GO 용어를 미리 설정된 숫자 조합인 GO 코드로 변환하는 GO 코드 변환부;

상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하고, 계산된 유사거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하며, 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 용어 모두에 대해 상기 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 생물학적 의미 추출부를 포함하는 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 2】

제1항에 있어서,

상기 GO 용어 할당부는 생물학 데이터베이스를 마이닝한 결과로부터 유전자에 상응하는 GO 용어를 할당하는 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 3】

제1항에 있어서,

상기 GO 코드 변환부는 변환하려는 GO 용어가 속하는 레벨, 변환하려는 GO 용어의 모(母)노드 정보 및 변환하려는 GO 용어가 속하는 레벨에서의 순서에 따라 GO 용어를 GO 코드로 변환하는 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 4】

제1항에 있어서,

상기 생물학적 의미 추출부는,

상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 최적 교차점들을 추출하는 최적 교차점 추출부;

상기 최적 교차점 정보를 이용하여 상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 유사 거리들을 계산하는 유사 거리 계산부;

상기 유사 거리 계산부에서 계산된 유사 거리들의 평균 유사 거리를 계산하는 평균 유사 거리 계산부;

상기 유사 거리 계산부에서 계산된 유사 거리들 중 최대 유사 거리를 판단하는 최대 유사 거리 판단부;

상기 미리 설정된 그룹에 속하는 모든 GO 용어에 대한 평균 유사 거리 및 최대 유사 거리를 비교하여, 최소의 평균 유사 거리 또는 최소의 최대 유사 거리에 상응하는 GO 트리 구조상의 용어를 상기 클러스터에 대응하는 최적의 매칭 노드로 판단하는 최적 매칭 노드 판단부를 포함하는 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 5】

제4항에 있어서,

상기 미리 설정된 그룹의 속하는 GO 용어는 GO 트리 구조에 포함된 모든 용어인 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 6】

제4항에 있어서,

상기 미리 설정된 그룹의 속하는 GO 용어는 사용자가 선택한 GO 트리 구조 레벨에 상응하는 GO 용어인 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 7】

제4항에 있어서,

상기 최적 교차점 추출부는 GO 트리 구조에서 두 개의 GO 용어를 모두 하위 레벨에 포함하는 상위 GO 용어들 중 가장 하위 레벨에 속하는 GO 용어를 최적 교차점으로 판단하는 것을 특징으로 하는 바이오 칩 분석 시스템.

【청구항 8】

제1항에 있어서,

GO 트리 구조의 각 레벨에는 미리 설정된 가중치가 부여되어 있으며, 상기 유사 거리 계산부에서 계산되는 유사 거리는 두 GO 용어의 최적 교차점이 속하는 레벨의 가중치인 것을 특

징으로 하는 바이오 칩 분석 시스템.

【청구항 9】

바이오 칩을 분석하기 위한 방법에 있어서,

a) 상기 바이오 칩 실험 결과의 통계적 클러스터링 결과를 입력받아 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 단계;

b) 상기 유전자마다 할당된 GO 용어를 각각 미리 설정된 숫자 조합인 GO 코드로 변환하는 단계;

c) 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하는 단계;

d) 상기 단계(c)에서 계산된 유사 거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하는 단계; 및

e) 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 모두에 대해 상기 단계 (c) 및 (d)를 반복하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 단계를 포함하는 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 10】

제9항에 있어서,

상기 단계 a)는 생물학 데이터베이스를 마이닝한 결과로부터 유전자에 상응하는 GO 용어를 할당하는 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 11】

제9항에 있어서,

상기 단계 b)는 변환하려는 GO 용어가 속하는 레벨, 변환하려는 GO 용어의 모(母)노드 정보 및 변환하려는 GO 용어가 속하는 레벨에서의 순서에 따라 GO 용어를 GO 코드로 변환하는 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 12】

제9항에 있어서,

상기 미리 설정된 그룹의 속하는 GO 용어는 GO 트리 구조에 포함된 모든 GO 용어인 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 13】

제9항에 있어서,

상기 미리 설정된 그룹의 속하는 GO 용어는 사용자가 선택한 GO 트리 구조 레벨에 상응하는 GO 용어인 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 14】

제9항에 있어서,

상기 단계(c)는,

상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 최적 교차점들을 추출하는 단계; 및

상기 최적 교차점 정보를 이용하여 상기 GO 트리 구조상의 용어와 상기 클러스터에 포함되는 유전자들에 할당된 GO 용어들과의 유사 거리들을 계산하는 단계를 포함하는 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 15】

제9항에 있어서,

상기 단계(e)는,

최소의 평균 유사 거리 또는 최소의 최대 유사 거리에 상응하는 GO 트리 구조상의 용어를 상기 클러스터에 대응하는 최적의 매칭 노드로 판단하는 것을 특징으로 하는 바이오 칩 분석 방법.

【청구항 16】

제14항에 있어서,

상기 최적 교차점들을 추출하는 단계는,

GO 트리 구조에서 두 개의 GO 용어를 모두 하위 레벨에 포함하는 상위 GO 용어들 중 가장 하위 레벨에 속하는 GO 용어를 최적 교차점으로 판단하는 것임을 특징으로 하는 바이오 칩 분석 방법.

【청구항 17】

제14항에 있어서,

GO 트리 구조의 각 레벨에는 미리 설정된 가중치가 부여되어 있으며,

상기 유사 거리들을 계산하는 단계는 두 GO 용어의 최적 교차점이 속하는 레벨의 가중치를 유사 거리로 계산하는 것임을 특징으로 하는 바이오 칩 분석 방법.

【청구항 18】

바이오 칩 분석 방법을 실행하기 위하여 디지털 처리 장치에 의해 실행될 수 있는 명령어들의 프로그램이 유형적으로 구현되어 있으며, 디지털 처리 장치에 의해 판독할 수 있는 기록 매체에 있어서,

상기 바이오 칩 분석 방법은,

a) 상기 바이오 칩 실험 결과의 통계적 클러스터링 결과를 입력받아 각 클러스터에 속하는 유전자들마다 관계된 GO 용어를 할당하는 단계;

b) 상기 유전자마다 할당된 GO 용어를 각각 미리 설정된 숫자 조합인 GO 코드로 변환하는 단계;

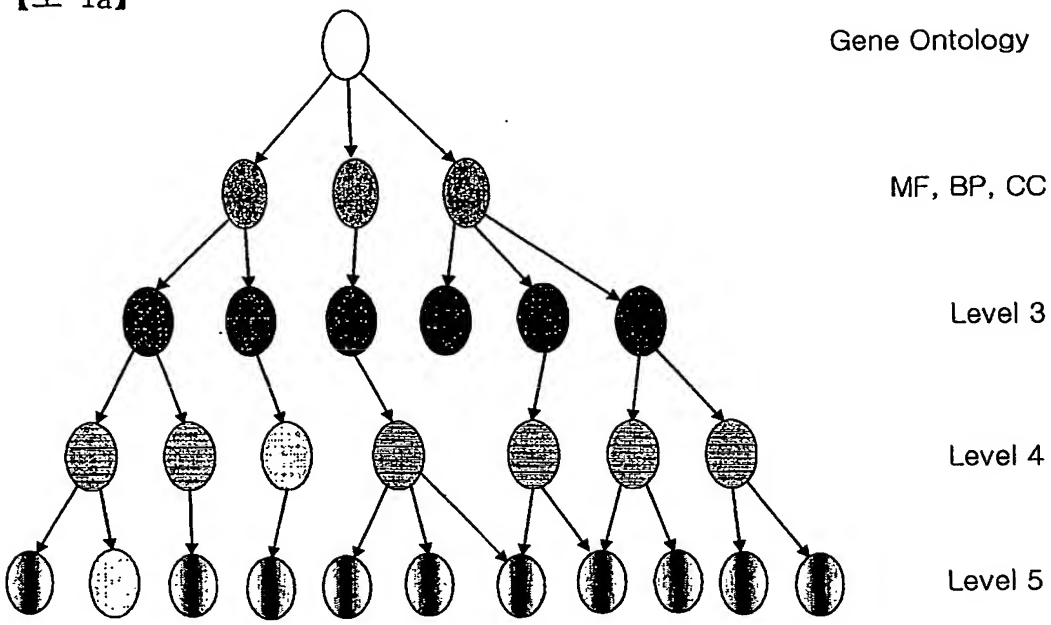
c) 상기 GO 코드를 이용하여, 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 중 하나와 상기 클러스터에 포함된 유전자들에 상응하는 GO 용어들과의 유사 거리를 계산하는 단계;

d) 상기 단계(c)에서 계산된 유사 거리들의 평균 유사 거리 및 최대 유사 거리 중 적어도 하나를 계산하는 단계; 및

e) 상기 미리 설정된 그룹에 속하는 GO 트리 구조상의 GO 용어 모두에 대해 상기 단계 (c) 및 (d)를 반복하여 상기 클러스터와 최적으로 매칭이 되는 GO 용어를 판단하는 단계를 포함하는 것을 특징으로 하는 기록 매체.

【도면】

【도 1a】



【도 1b】

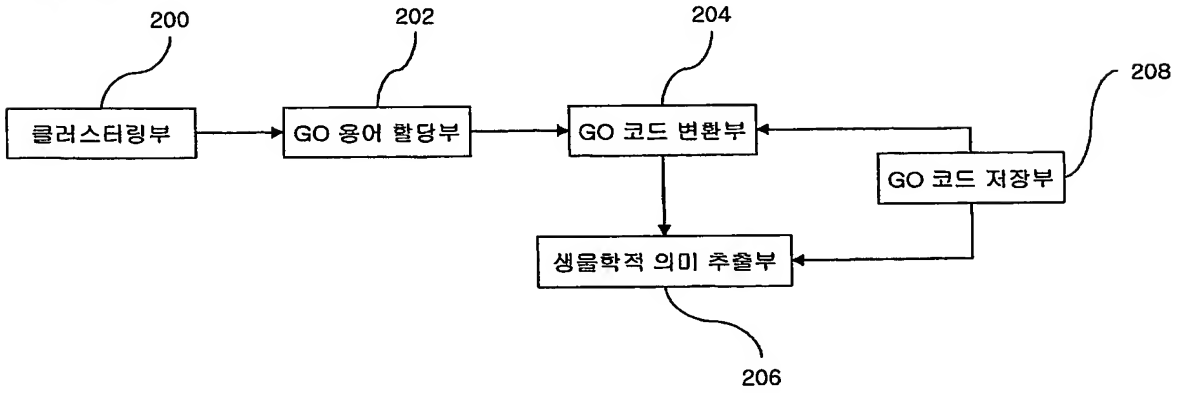
<Biological Process>

Gene_Ontology
- biological_process
  - death
    - cell death
    - apoptosis
      - anti-apoptosis +
      - apoptotic program +
      - induction of apoptosis +
      - killing of inflammatory cells
      - killing transformed cells
      - killing virus-infected cells
      - peripheral killing of activated T cells
    - non-apoptotic cell death

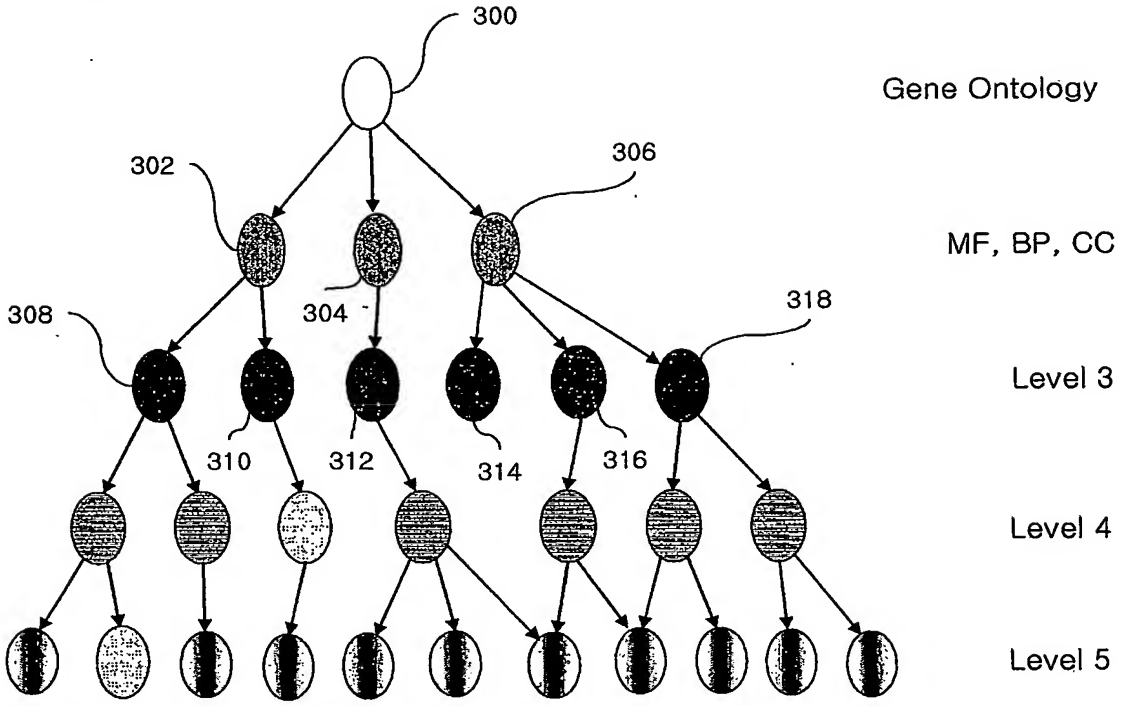
<Molecular Function>

Gene_Ontology
- molecular_function
  - apoptosis regulator
    - apoptosis activator
      - apoptotic protease activator
    - apoptosis inhibitor

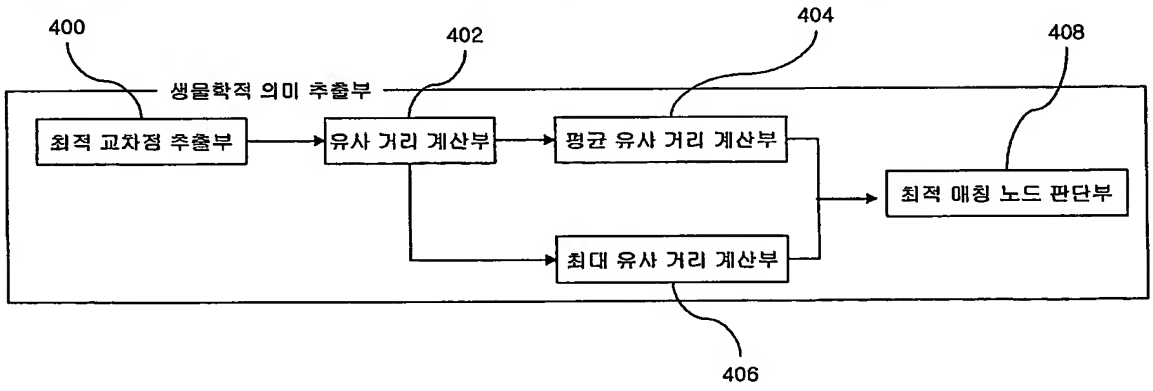
【도 2】



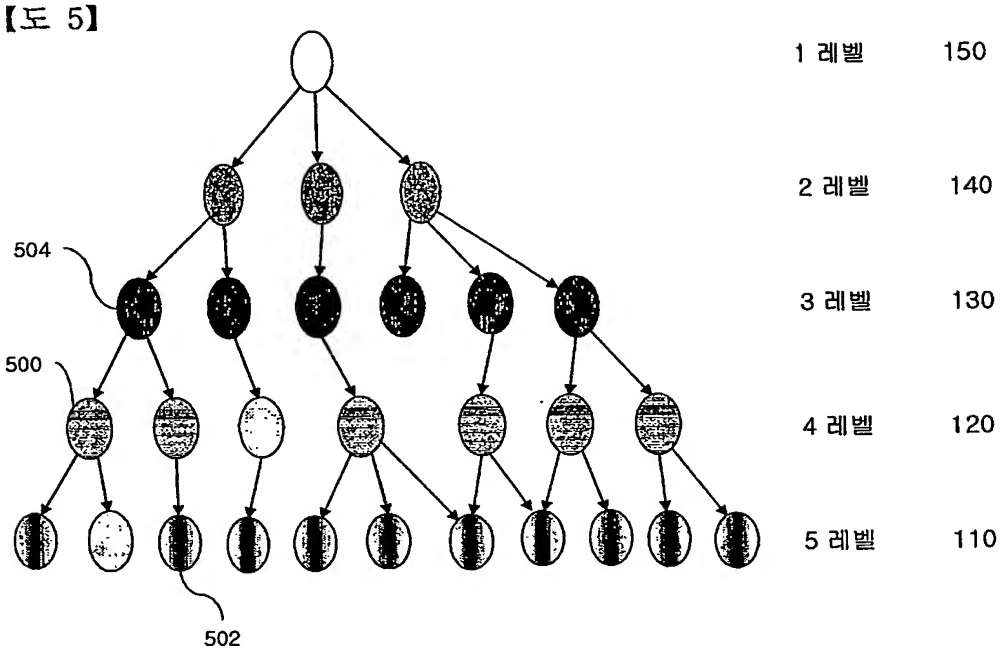
【도 3】


300
Gene Ontology
302
306
MF, BP, CC
308
304
318
Level 3
310
312
314
316
Level 4
Level 5

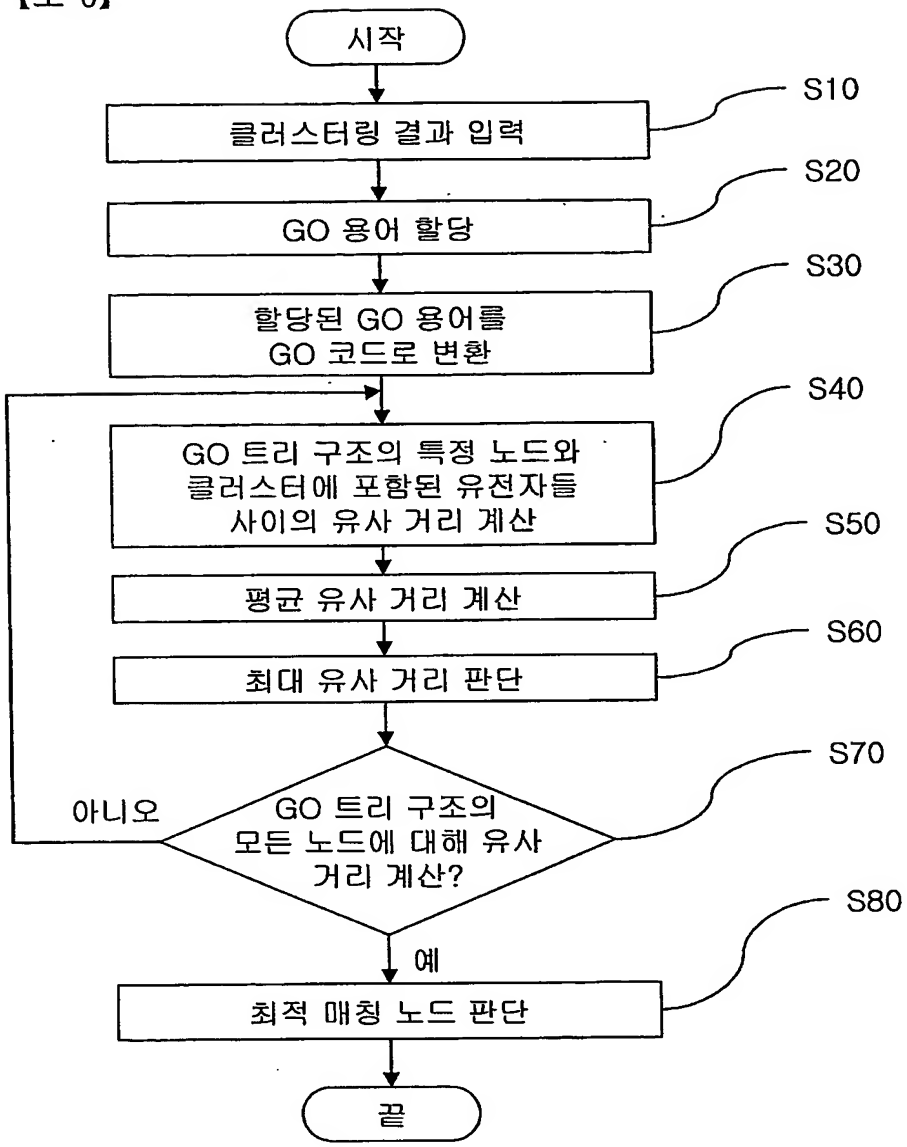
【도 4】


400
402
404
408
생물학적 의미 추출부
최적 교차점 추출부
유사 거리 계산부
평균 유사 거리 계산부
최적 매칭 노드 판단부
최대 유사 거리 계산부
406

【도 5】



【도 6】


시작
클러스터링 결과 입력
S10
GO 용어 할당
S20
할당된 GO 용어를
GO 코드로 변환
S30
GO 트리 구조의 특정 노드와
클러스터에 포함된 유전자들
사이의 유사 거리 계산
S40
평균 유사 거리 계산
S50
최대 유사 거리 판단
S60
GO 트리 구조의
모든 노드에 대해 유사
거리 계산?
S70
아니오
예
최적 매칭 노드 판단
S80
끝